



Earth Science Data Systems
Interagency Implementation and Advanced Concepts Team

A collective agenda on AI for Earth sciences

Manil Maskey, Ph.D.

AI for Good
February 16, 2022

What I hope to do today

Data systems perspective on AI for Earth science

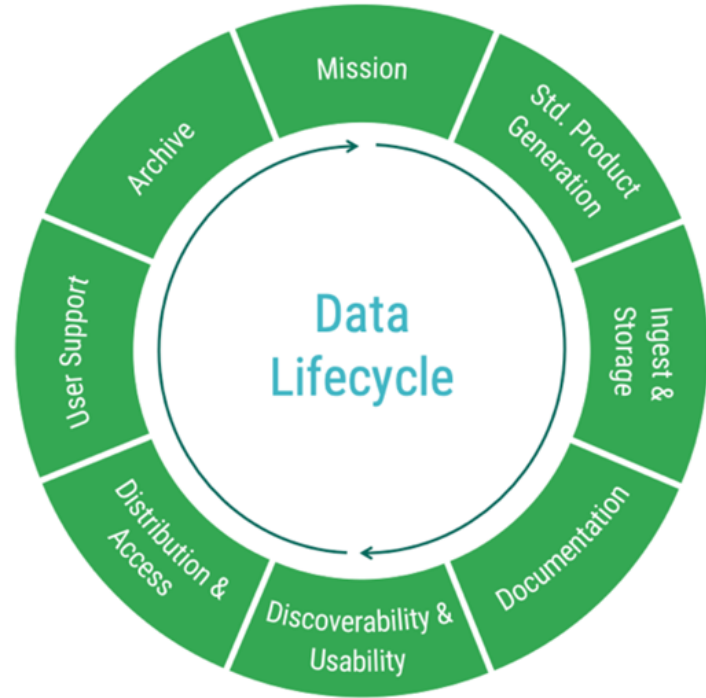
- AI for core data systems services
 - Search
 - Knowledge discovery
- Enabling AI to advance Earth science
 - Data (labeled training data) is the proprietary differentiator
 - Transitioning AI models to Production
 - Citizen science



NASA's Earth Science Data Systems Program

Single largest repository of Earth Science Data

Manages NASA's Earth science data through the entire data life cycle



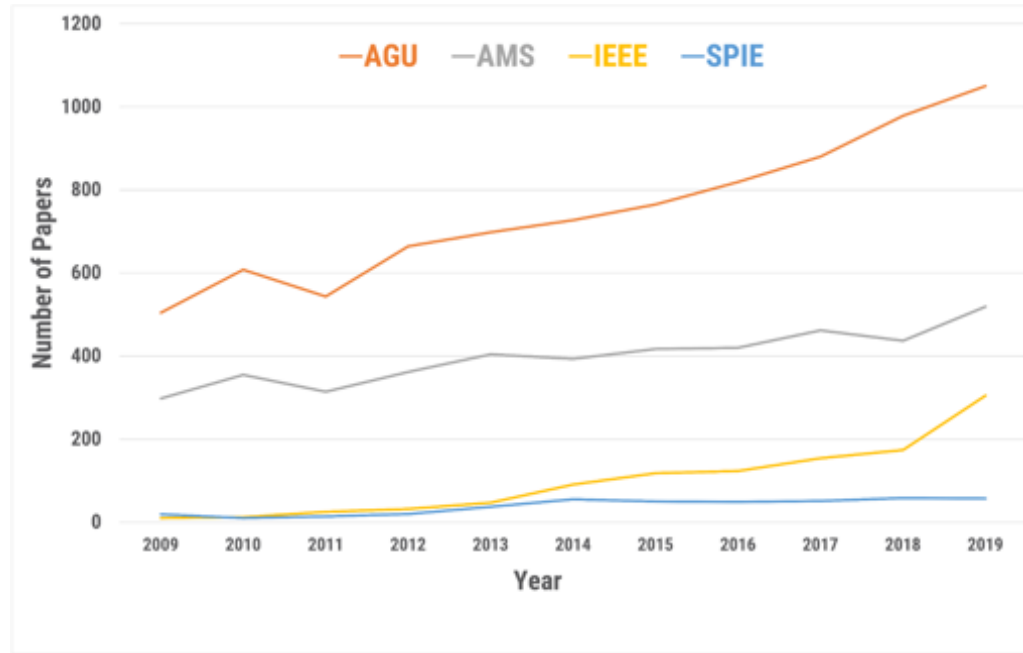
AGU topic trend

Number of journal articles per keyword by year



2009

Publication trend - ML in Earth sciences



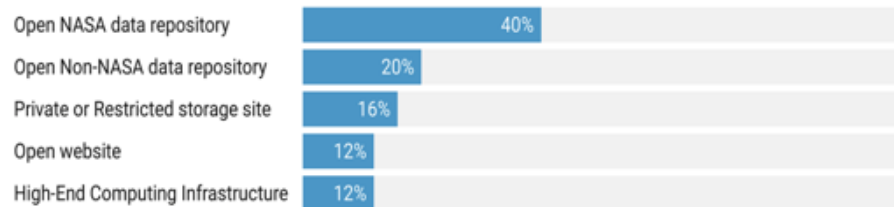
Rapid adoption of AI/ML by Earth science researchers

Virts et al. (2020)

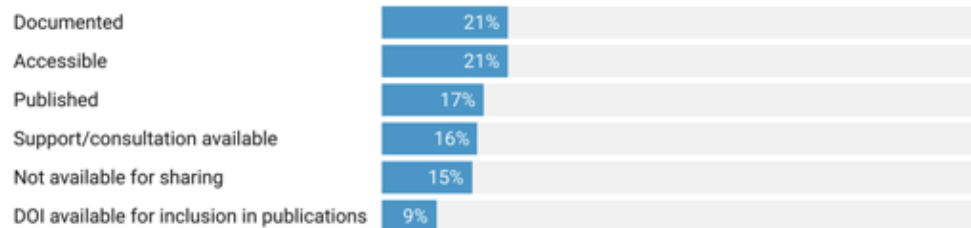
Maskey et al. (2020)

NASA Science survey - AI and data

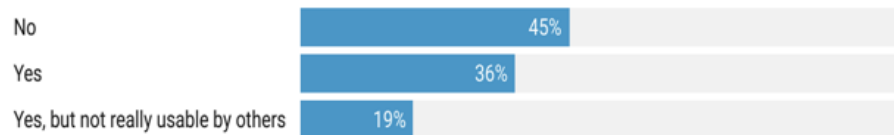
Source of data used



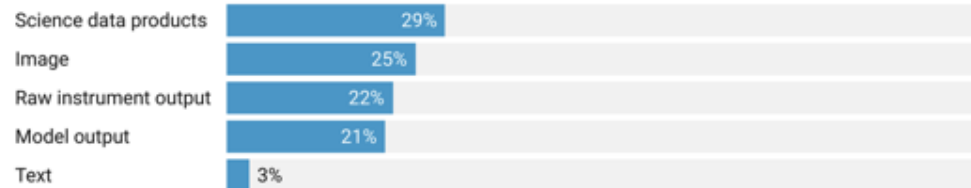
How re-usable is your training data?



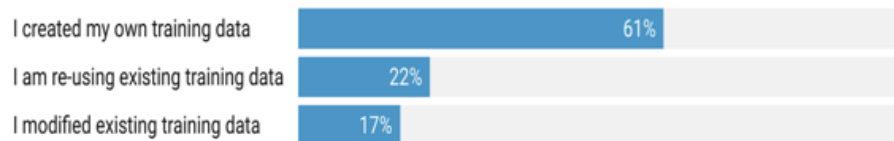
Is there a catalog of training data for your use?



What type of data do you use for AI?



How did you construct training data?

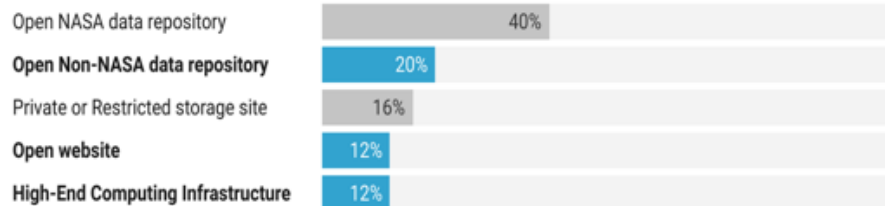


Amount of effort required to prepare data for AI?

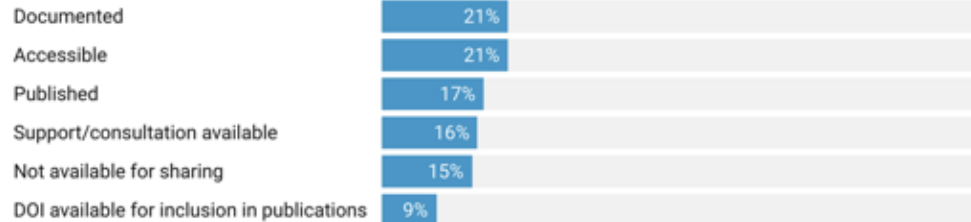


NASA Science survey - AI and data

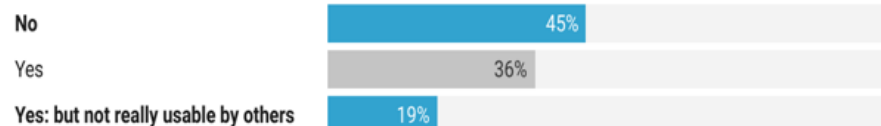
Source of data used



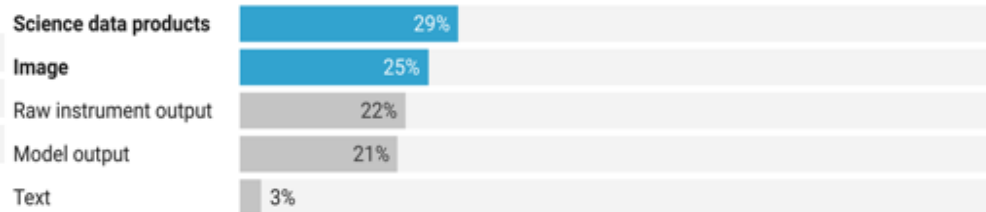
How re-usable is your training data?



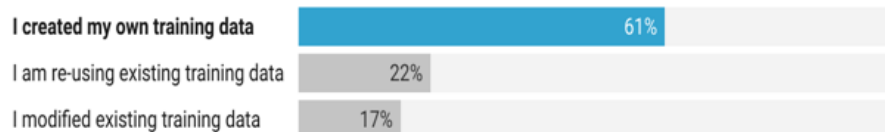
Is there a catalog of training data for your use?



What type of data do you use for AI?



How did you construct training data?



Amount of effort required to prepare data for AI?



Maximizing Knowledge Discovery

Why?

Increasing Earth science data archives require non-traditional approaches to data management

Data driven technologies (AI) to provide advanced search capabilities

Machine learning-based approach - provide automated detection of Earth science events from image archives

Catalog of events can provide a novel way to explore large archives of data

Discover and explore Earth science data archives around events using machine learning (ML) techniques

WORLDVIEW

Layers Events Data

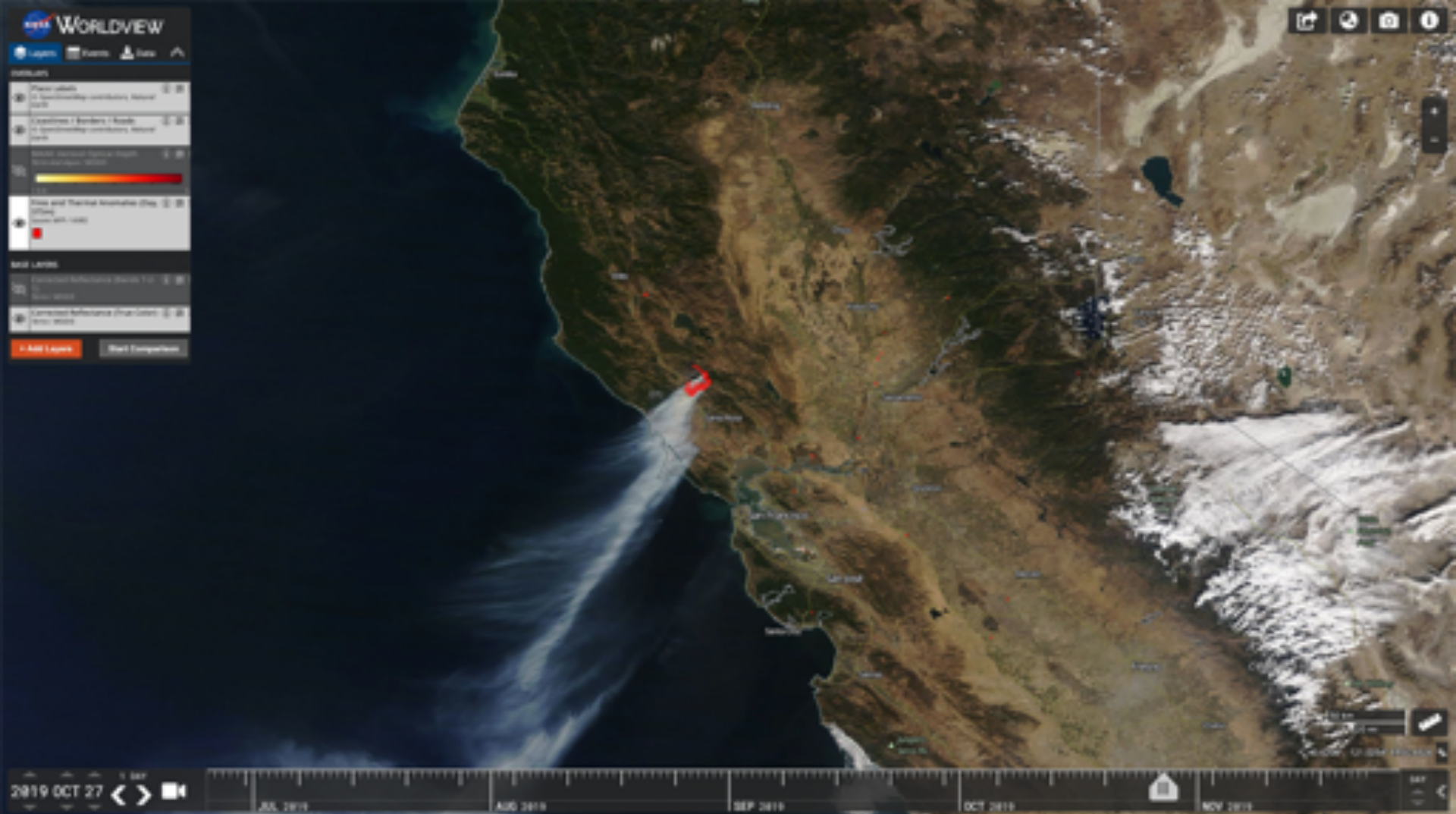
CONTOURS

- Phenology (US Department of Agriculture, National Forest Service)
- Vegetation - Boundaries - Roads (US Department of Agriculture, National Forest Service)
- Global Wetland Inventory (Global Wetland Inventory)
- Fire and Thermal Anomalies (Day Offsets) (Global Wetland Inventory)

SOIL LAYERS

- Global Soil Database (Global Soil Database)
- Global Soil Database (Global Soil Database)

[Add Layers](#) [Start Comparison](#)





Sort by: ☐ Relevance ☒ Only include collections with granules ☐ Include non-ECOSIS collections

Tip: Add a collection to your cart to compare and download data sets



MODIS/Terra Thermal Anomalies/Fire 5-Day L3 Global V05 SW SW 1000

5 Granules • 2000-05-18 ending • The Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Thermal anomalies and fire 5-day (MT150005) Version 5 data are generated at 1 kilometer (km) spatial resolution as a Level 3 product. The MT150005 granule composite contains maximum value of the individual fire pixel values detected during the eight days of acquisition. The Science Dataset (SDS) layers include the fire mask, and quality indicators. Improvements/Changes from Previous Versions: "...

[View metadata](#)



MODIS/Terra Thermal Anomalies/Fire Daily L3 Global V05 SW SW 1000

5 Granules • 2000-05-18 ending • The Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Thermal anomalies and fire daily (MT150005) Version 5 data are generated every eight days at 1 kilometer (km) spatial resolution as a Level 3 product. The MT150005 granule composite contains eight consecutive days of the data consistently packaged into a single file. The Science Dataset (SDS) layers include the fire mask, and quality indicators, maximum the relative power (RTP), and the position of the fire spot within the scene...

[View metadata](#)



MODIS/Aqua Thermal Anomalies/Fire Daily L3 Global V05 SW SW 1000

5 Granules • 2000-05-18 ending • The Aqua Moderate Resolution Imaging Spectroradiometer (MODIS) Thermal anomalies and fire daily (MT150005) Version 5 data are generated every eight days at 1 kilometer (km) spatial resolution as a Level 3 product. The MT150005 granule composite contains eight consecutive days of the data consistently packaged into a single file. The Science Dataset (SDS) layers include the fire mask, and quality indicators, maximum the relative power (RTP), and the position of the fire spot within the scene...

[View metadata](#)

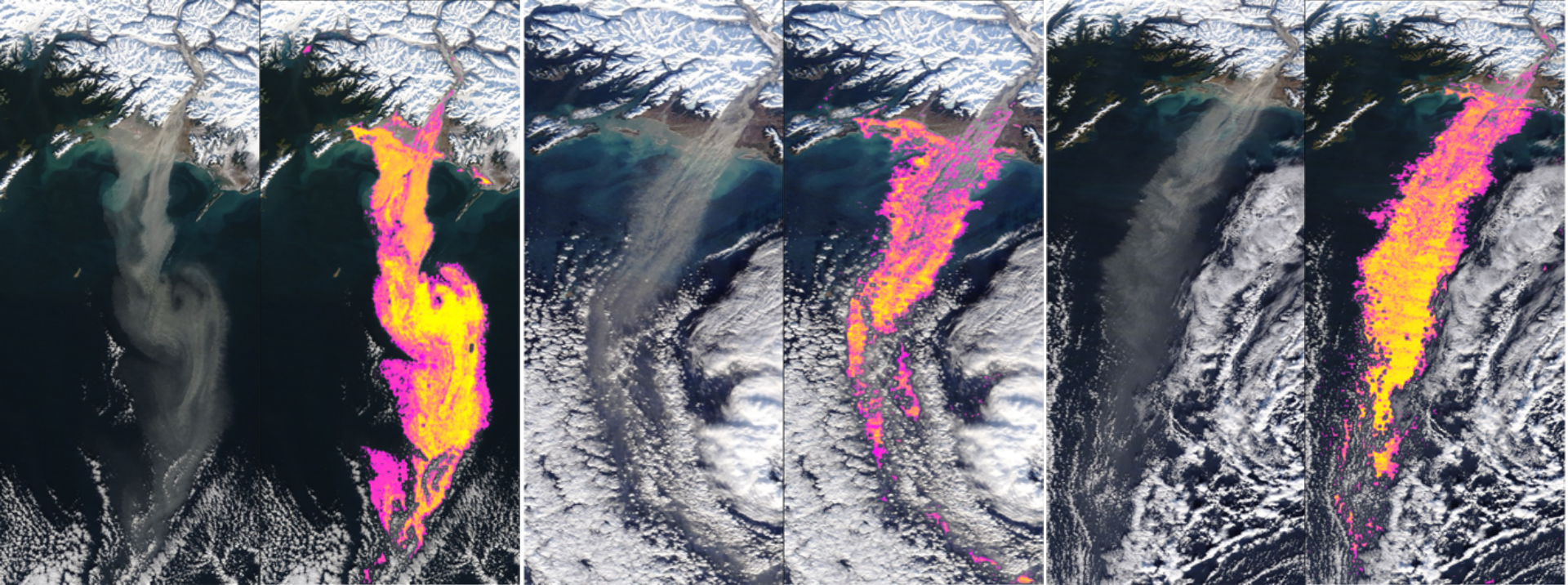


MODIS/Aqua Thermal Anomalies/Fire 5-Day L3 Global V05 SW SW 1000

5 Granules • 2000-05-18 ending • The Aqua Moderate Resolution Imaging Spectroradiometer (MODIS) Thermal anomalies and fire 5-day (MT150005) Version 5 data are generated at 1 kilometer (km) spatial resolution as a Level 3 product. The MT150005 granule composite contains maximum value of individual fire pixel values detected during the eight days of acquisition. The Science Dataset (SDS) layers include the fire mask, and quality indicators. Improvements/Changes from Previous Versions: "...

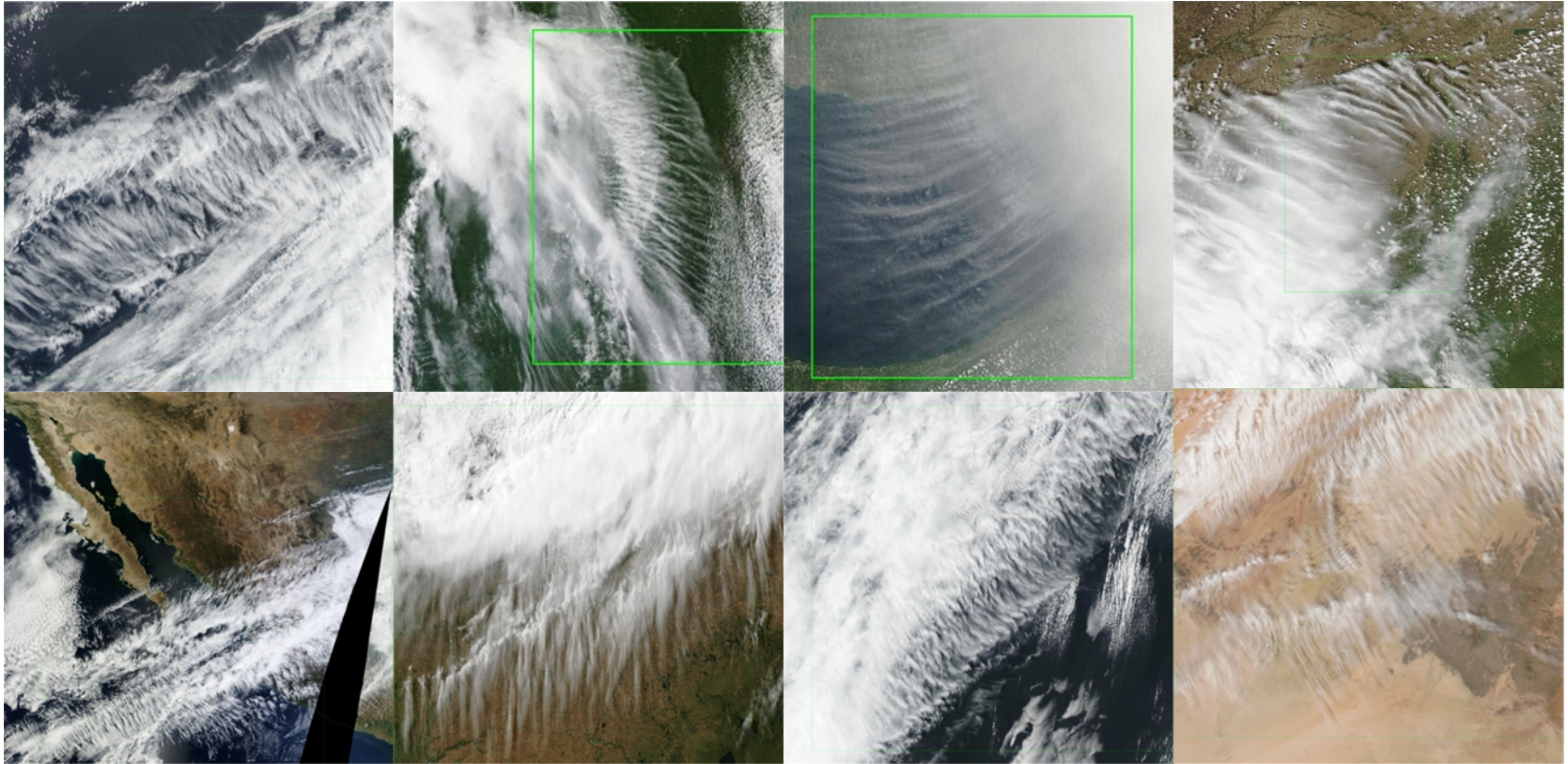
[View metadata](#)





High latitude dust

Transverse cirrus bands



Welcome to the Phenomena Detection Portal

We are using machine learning for real-time
detection of Earth science phenomena.

Types

03

so far

Detections

98,627

and counting

Confidence score

89.61%

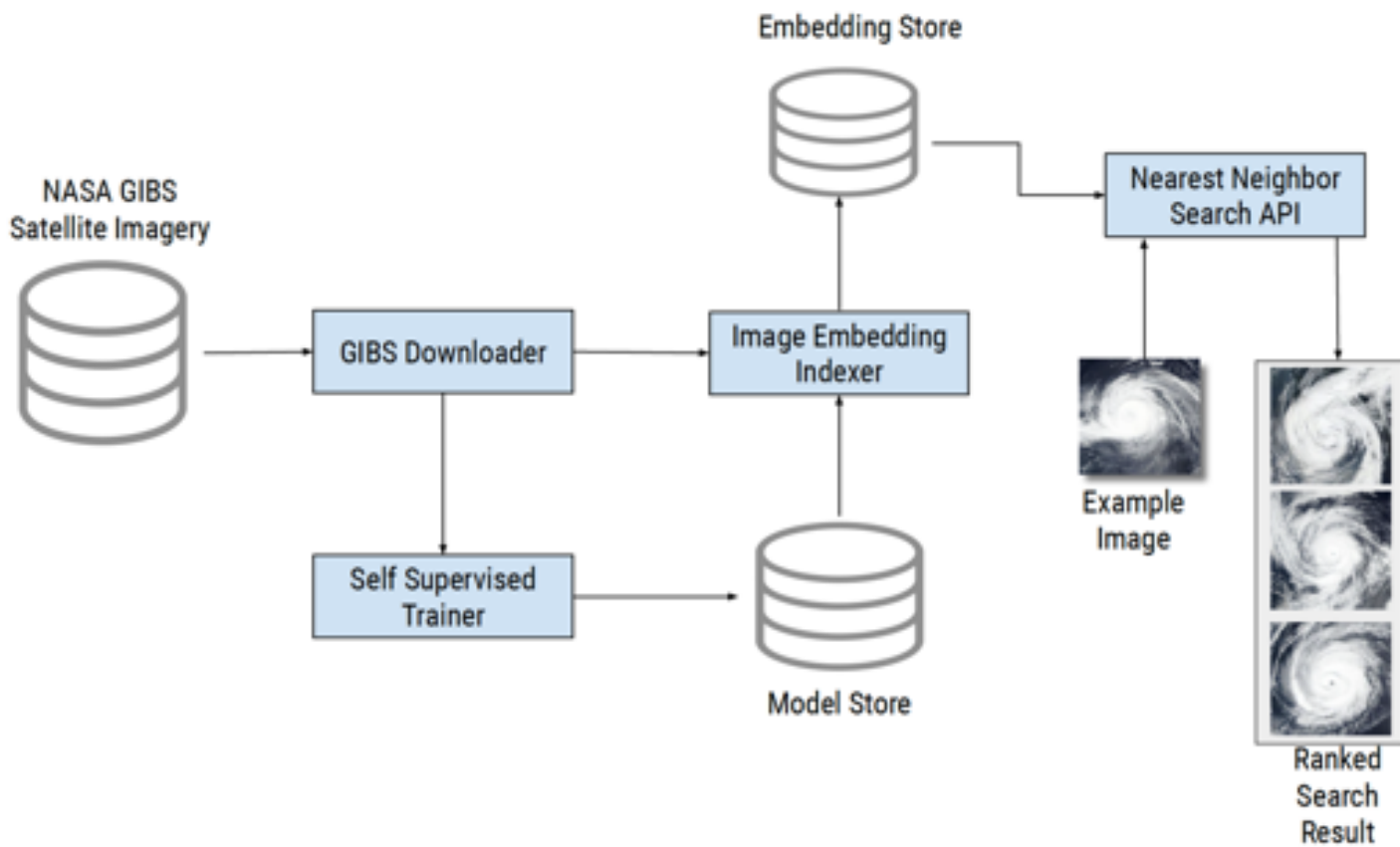
on average

Start exploring

Learn more



Search by Example



Similarity Search Demo UI

CLICK THE MAP TO SELECT AN INPUT IMAGE / WIKEDRY DATE 1/20/22 / IMAGE NEIGHBORS 1



Results



2019-07-11

Scaled
SSL

Augment data stewardship processes

Why?

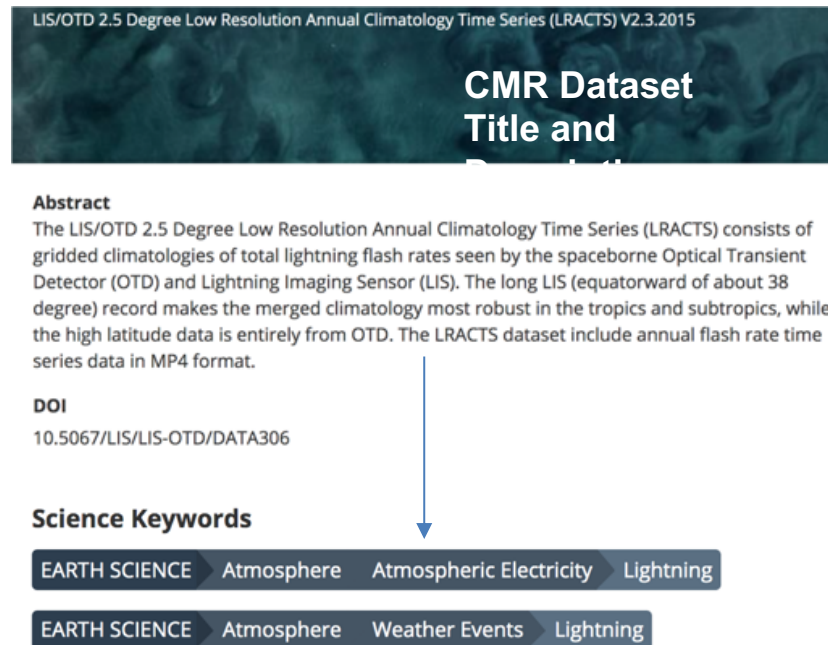
Assigning science keywords is currently a manual process, which is prone to human error and inconsistencies.

Metadata managed across a network of multiple data centers (i.e. keywords not assigned by a central entity)

Keywords may be assigned by non-subject matter experts (SMEs)

Improve metadata quality

Provide objective and consistent approach to keyword assignment



LIS/OTD 2.5 Degree Low Resolution Annual Climatology Time Series (LRACTS) V2.3.2015

CMR Dataset Title and

Abstract
The LIS/OTD 2.5 Degree Low Resolution Annual Climatology Time Series (LRACTS) consists of gridded climatologies of total lightning flash rates seen by the spaceborne Optical Transient Detector (OTD) and Lightning Imaging Sensor (LIS). The long LIS (equatorward of about 38 degree) record makes the merged climatology most robust in the tropics and subtropics, while the high latitude data is entirely from OTD. The LRACTS dataset include annual flash rate time series data in MP4 format.

DOI
10.5067/LIS/LIS-OTD/DATA306

Science Keywords

EARTH SCIENCE > Atmosphere > Atmospheric Electricity > Lightning

EARTH SCIENCE > Atmosphere > Weather Events > Lightning

Approach – build word embeddings

Journal Name	Date Published																	
	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Atmospheric Science Letters		5	23	34	27	22	42	39	59	64	34	80	69	67	69	65	67	58
Earth and Space Science													1	24	28	25	42	88
Earth's Future												13	26	24	52	58	80	56
Eos, Transactions American Geophysical Union										46	37			1				
Geochemistry, Geophysics, Geosystems	88	267	266	573	155	143	208	314	164	203	192	247	196	154	185	167	64	29
Geoscientific Data Discovery																22	36	54
Geophysical Research Letters	5,155	1,436	1,558	1,698	1,700	1,533	1,559	1,290	1,005	1,234	1,058	1,354	1,200	1,388	1,491	1,390	1,507	1,136
Global Biogeochemical Cycles	132	136	124	120	75	66	94	66	83	64	96	76	72	76	75	74	73	71
Journal of Advances in Modeling Earth Systems								6	3	9	28	35	49	88	66	113	125	134
Journal of Geophysical Research							56				20							
Journal of Geophysical Research: Atmospheres	873	1,288	186	706	784	965	972	737	954	949	782	969	815	788	841	758	784	538
Journal of Geophysical Research: Biogeosciences				73	79	140	391	312	146	185	130	92	109	138	138	132	45	24
Journal of Geophysical Research: Earth Surface		12	47	95	84	145	130	113	134	132	141	117	92	92	93	59	44	30
Journal of Geophysical Research: Oceans	253	492	375	314	317	341	434	323	382	501	418	338	345	330	325	388	492	284
Journal of Geophysical Research: Planets	137	279	175	126	162	150	795	130	147	171	172	107	89	78	75	103	142	92
Journal of Geophysical Research: Solid Earth	345	683	365	319	377	435	436	376	508	450	378	297	316	307	314	388	126	96
Journal of Geophysical Research: Space Physics	409	563	498	525	475	447	589	533	756	614	694	490	503	541	592	542	466	446
Meteorological Applications								7	47	69	76	76	46	61	67	76	1	2
Paleogeography	65	189	98	82	67	61	85	47	64	59	55	45	59	76	56	58		
Paleogeography and Paleoclimatology																	24	26
Quarterly Journal of the Royal Meteorological Society									6	185		203	168	197	178			
Radio Science	188	132	146	114	94	122	91	108	59	136	79	52	69	70	309	93	94	51
Reviews of Geophysics	9	23	12	11	8	14					12	12	14	23	16	22	36	17
Space Weather		16	57	53	48	47	44	48	48	55	63	65	53	47	68	88	125	58
Tectonics	55	45	79	88	58	73	66	60	73	67	78	58	38	83	99	119	41	58
Water Resources Research	307	386	319	337	378	364	414	356	350	403	408	447	412	365	403	397	453	346

88,410

documents

530 million

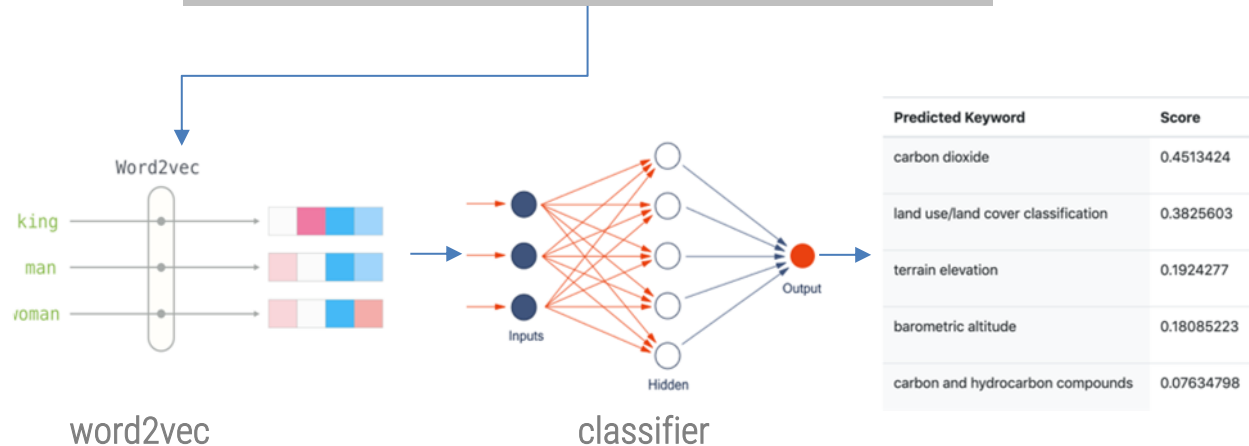
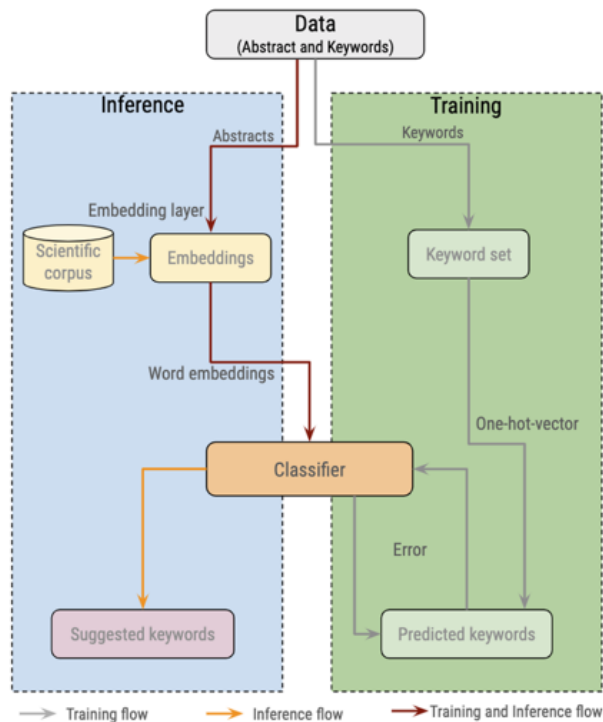
words

5.5 million

unique words

Automated keyword assignment




Version 7.3 is the current version of the data set. Version 3.5 is no longer available and has been superseded by Version 7.3. This data set is currently provided by the OCO (Orbiting Carbon Observatory) Project. In expectation of the OCO-2 launch, the algorithm was developed by the Atmospheric CO2 Observations from Space (ACOS) Task as a preparatory project, using GOSAT TANSO-FTS spectra. After the OCO-2 launch, "ACOS" data are still produced and improved, using approaches applied to the OCO-2 spectra. The "ACOS" data set contains Carbon Dioxide (CO2) column averaged dry air mole fraction for all soundings for which retrieval was attempted. These are the highest-level products made available by the OCO Project, using TANSO-FTS spectral radiances, and algorithm build version 7.3. The GOSAT team at JAXA produces GOSAT TANSO-FTS Level 1B



Keyword recommender


EARTHDATA Powered by EODORS

[Collections](#) [Custom](#)

Int. Model for Classification  Concept ID  Version 

Word2Vec

[Suggest Keywords](#)

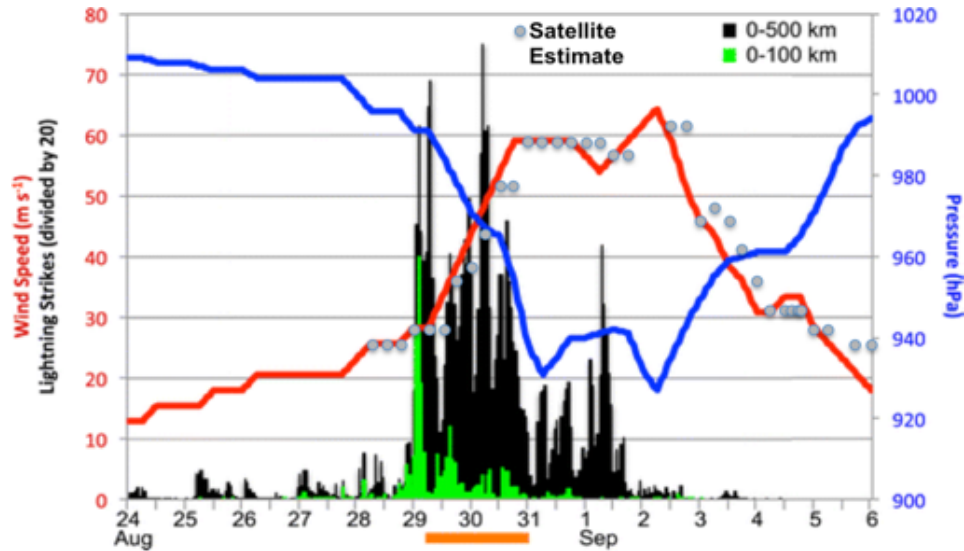


[NASA Official](#) [NASA Remuneration](#) | [Web Privacy Policy](#) [Data](#) [Information Policy](#) [Communications Policy](#) [Freedom of Information Act](#) [USA.gov](#)

Hurricane intensity estimation system

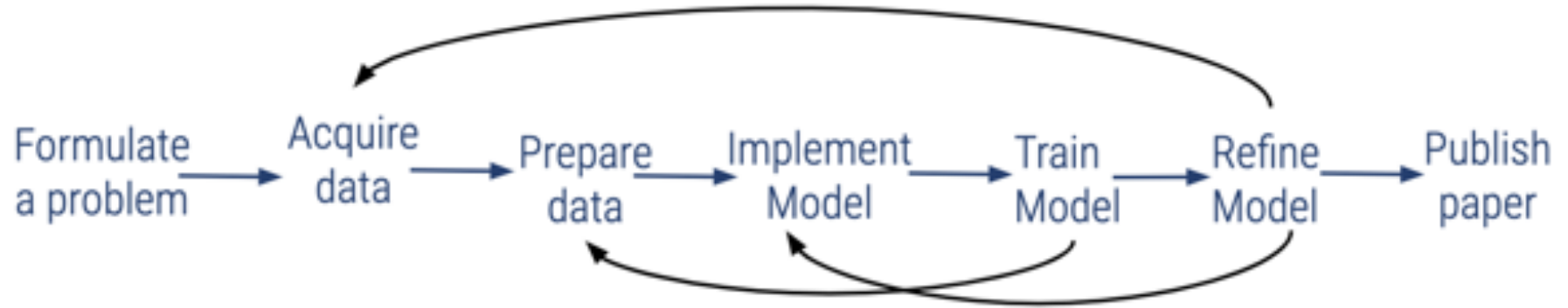
AI and satellite imagery to estimate hurricane wind speed

Hurricane Earl, 2010



Adapted from Stevenson et al. (2014). Time series of satellite-derived intensity estimates (circles) for Hurricane Earl (2010), added to best track intensities and lightning flash rate time series.

ML in literature



We have a model....now what?

Going extra mile

Interpretability + model inspection

Interpret prediction data – prediction output maybe just numbers

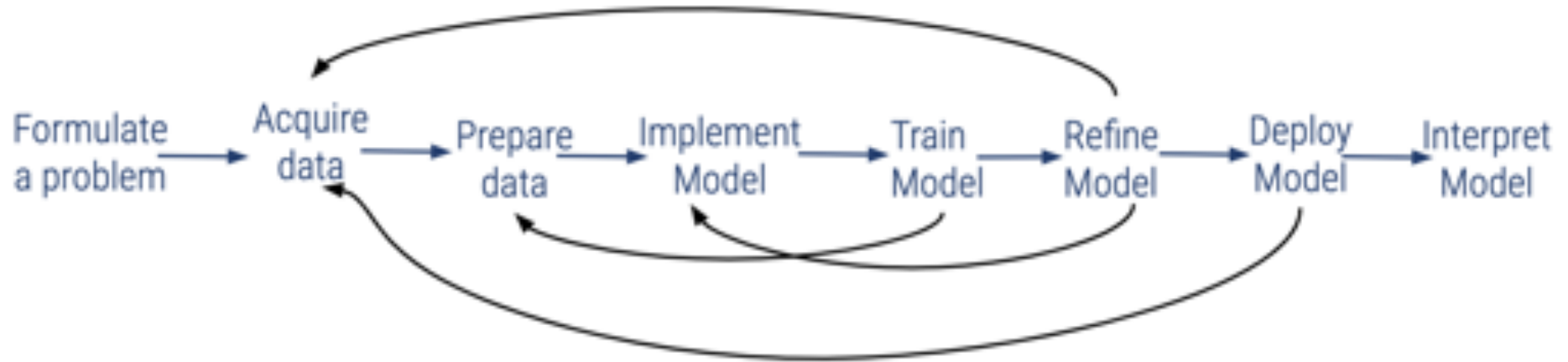
Questions:

Does the model confidence remain the same over time?

How do you maintain?

How do you complete the loop with new training data?

ML lifecycle - iterative



Deployment to production

Performance requirements

Metrics and baselines with initial models

Monitor over time

Back-testing

Model and software will change

Testing model changes on historical data

Run current production model to baseline performance

Run new models, competing for production

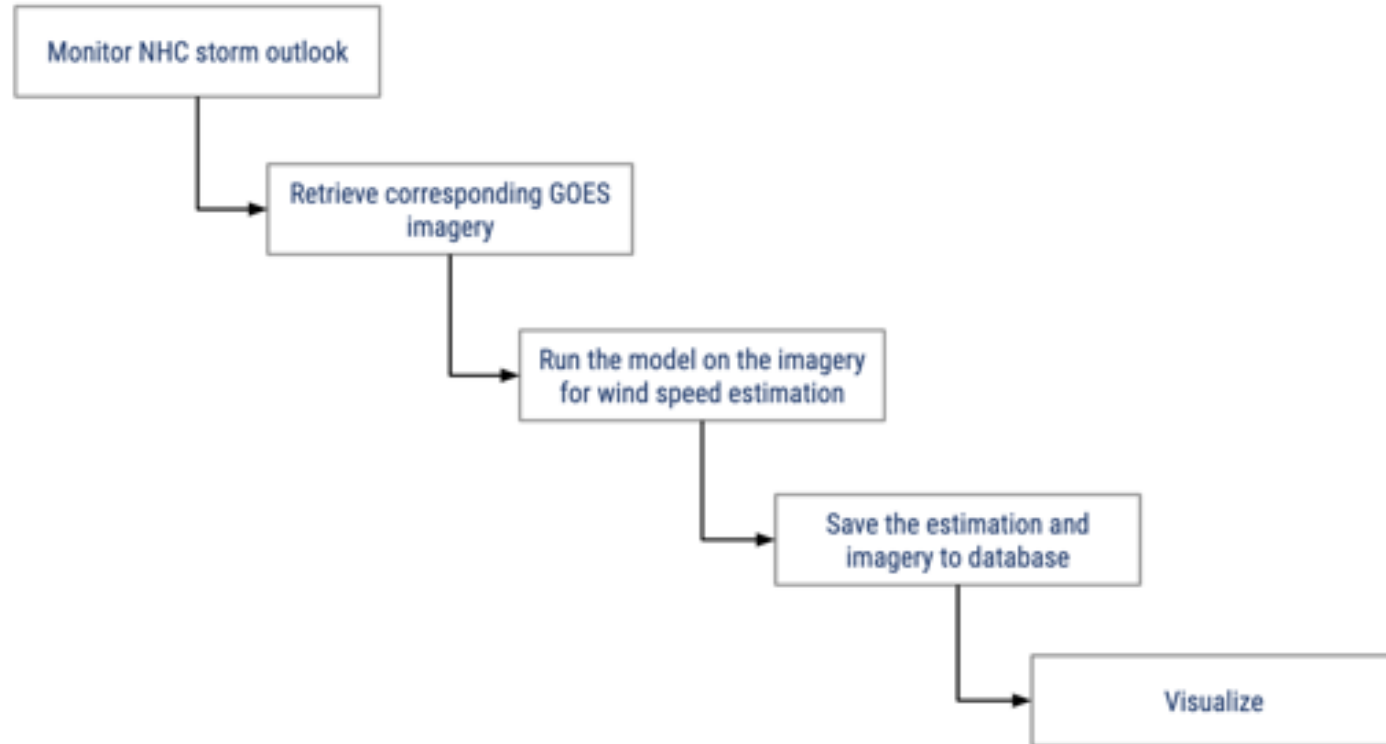
Now-testing

Testing of production model on latest data

Can we get early warning that the model may be faltering?

- Content drift: training data exploited by model are subtly changing with time

Workflow



<https://impact.earthdata.nasa.gov/hurricane/>

Deep Learning-based Hurricane Intensity Estimator

Applying machine learning to objectively estimate tropical cyclone intensity.

Explore

or

Read more



Dark: show again

Using community to advance model development

Data science competitions

- Benchmark datasets and challenge problems have played an important role in driving progress in AI
 - Enables rigorous performance comparison
- Foster the learning of best practices
- Stimulate the abilities in problem-solving
- Encourage creativity and group work
- Give learners the chance to interact with new platforms and algorithms
- Citizen science

Data science competition

“Wind-dependent Variables: Predict Wind Speeds of Tropical Storms”

- Leverage community to enhance solution to existing problem using open data
- Test whether high-quality datasets produces better models via open competition

- **Industry** partnership
- **733** participants
- **2756** entries

Wind-dependent Variables: Predict Wind Speeds of Tropical Storms
HOSTED BY RADIANT EARTH FOUNDATION

Prizes show that L1-Neural Networks can capture key patterns in the satellite imagery of storms to estimate wind speed, we seek to improve the accuracy for operational applications.

Task


The goal of this challenge is to estimate the wind speeds of storms at different points in time using satellite images captured throughout a storm's life cycle and the temporal memory of the storm. Radiant Earth Foundation has worked with the NASA IMPACT team to assemble a data set of tropical storm imagery which includes single-band satellite images at a long wave infrared frequency and corresponding wind speed estimations. Improving total and gust estimates from satellite imagery could mean significant improvements in short-term storm intensity forecasting, risk assessment models, and disaster readiness and response.

If the winning solution of this competition performs better than the existing model running on *Hurricane Intensity Estimates*, the model will be replaced with credit given to the winner.

Place	Prize Amount
1st	\$4000 and \$6000 Azure credit
2nd	\$4000 and \$6000 Azure credit
3rd	\$4000 and \$6000 Azure credit


Competition End Date:
Feb. 2, 2021, 11:59 p.m. UTC

This challenge is convened by our friends at Radiant Earth Foundation.



 **Radiant Earth Foundation**
EARTH IMAGERY FOR IMPACT

With generous support from:



CONVENING SPONSOR


NASA Earth Science Data Systems Program

GOLD SPONSORS


SILVER SPONSORS

AZURE CREDIT SPONSOR



TECHNICAL SUPPORTER



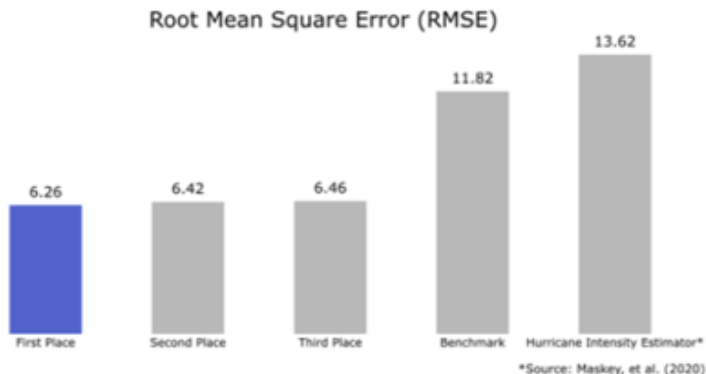
Banner image courtesy of the Centers for Disease Control and Prevention.

Data science competition

The Results

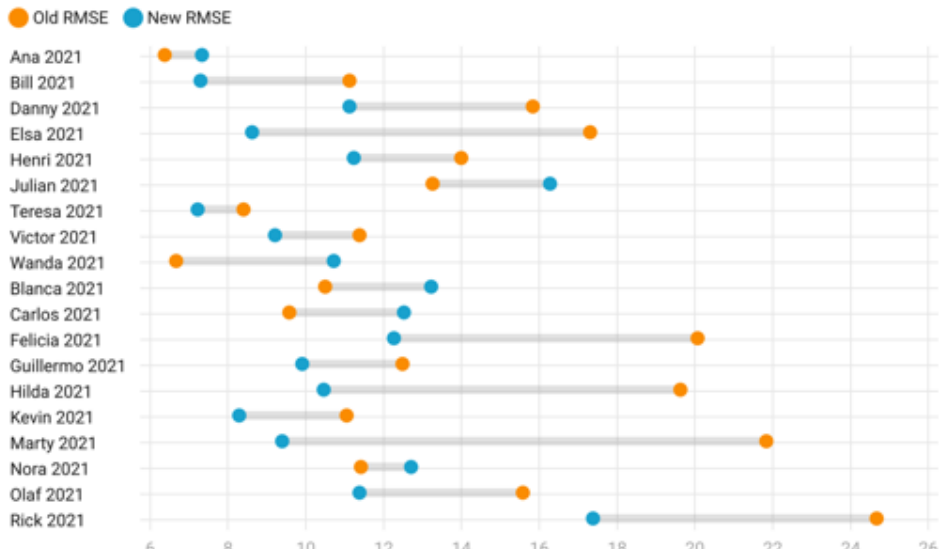
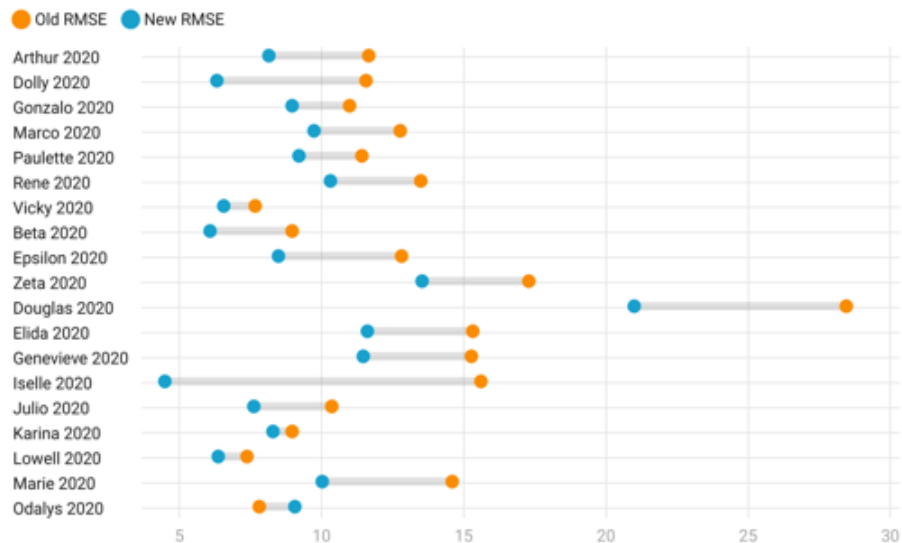
Over 700 participants stepped up to this important challenge, generating more than 2,700 entries.

Each of the top three models achieved at least a 50% reduction in Root Mean Square Error (lower is better) as compared to the existing model!



Winning solutions were able to take advantage of the relative timing of images in a storm sequence to produce targeted wind speed estimates based on temporal trends. As a result, these solutions can help to improve disaster readiness and response efforts around the world by equipping response teams with more accurate and timely wind speed measurements. All of the prize-winning solutions from this competition are linked below and made available for anyone to use and learn from.

New Model – in depth analysis



New Model – in depth analysis

● Old RMSE ● New RMSE



● Old RMSE ● New RMSE



Flood extent detection

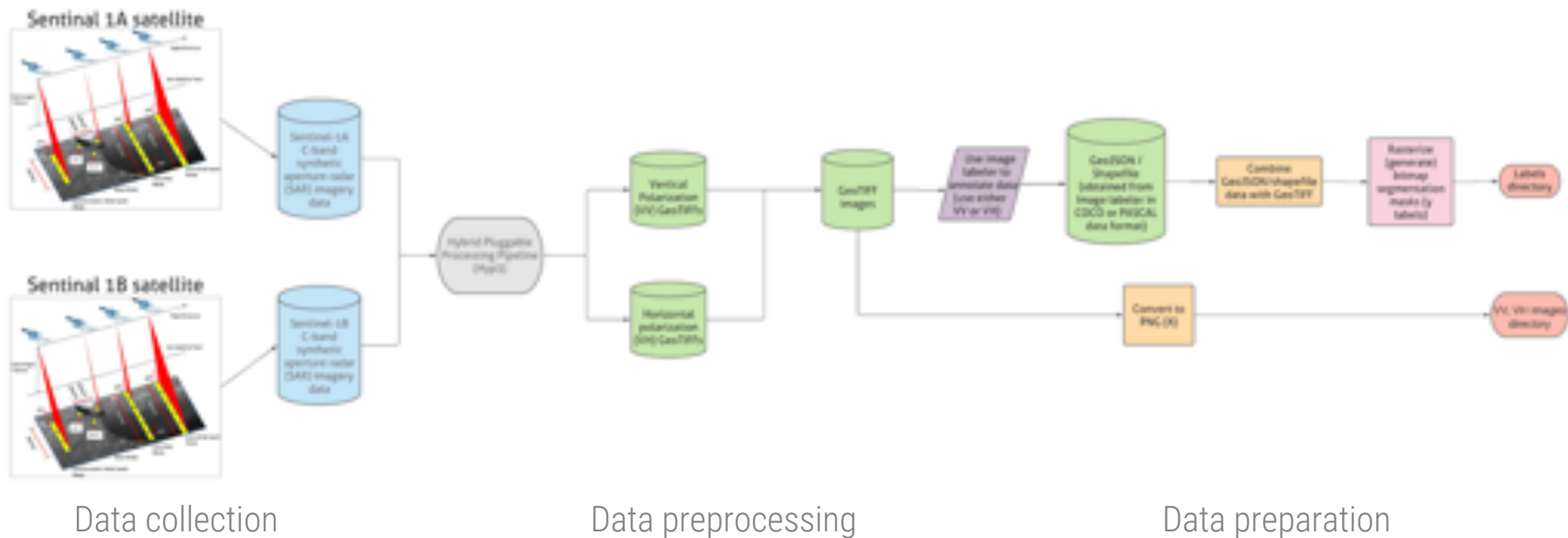
Flood extent detection

- A major natural disaster
- Widespread damage – property, agriculture
- Displacement, insurance, long-term socio-economic consequences
- Causes:
 - Persistent rainfall
 - Severe storm
 - High-tides
 - Storm surge from cyclones

Problem

- Detecting flood extent is difficult
- Monitoring extent of flood events in-situ – hazardous to operate in a disaster zone
- Potential solution:
 - Remote sensing in conjunction with ML has been used in the community to monitor these events
- Need:
 - Large amounts of clean and labeled data

Data acquisition



Data labeling

- 6 Atmospheric science/Earth science students
- 2 Domain scientists
- Training sessions
- Validation

<https://impact.earthdata.nasa.gov/labeler/>

ImageLabeler

WELCOME TO IMAGE LABELER

**A web application
to create and manage
labeled Earth science images
for machine learning.**

[Sign Up](#)

[Sign In](#)

© 2018 NASA

Labeled data

~66k images (33k VV + 33k VH images including swath gap artifacts)

Native resolution : 5x20m

Train (24300):

- Nebraska (1741 sq. km.) (~43 %)
- North Alabama (13789 sq. km.) (~43%)
- Bangladesh (7150 sq. km.) (~13%)

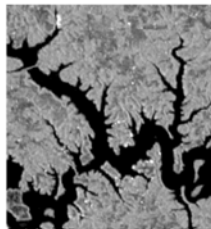
Validation (6500):

- Florence (7197 sq. km.)

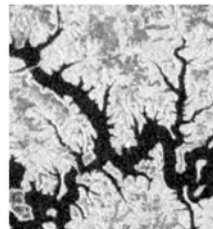
Test (1600):

- Red River North (6746 sq. km.)

VV

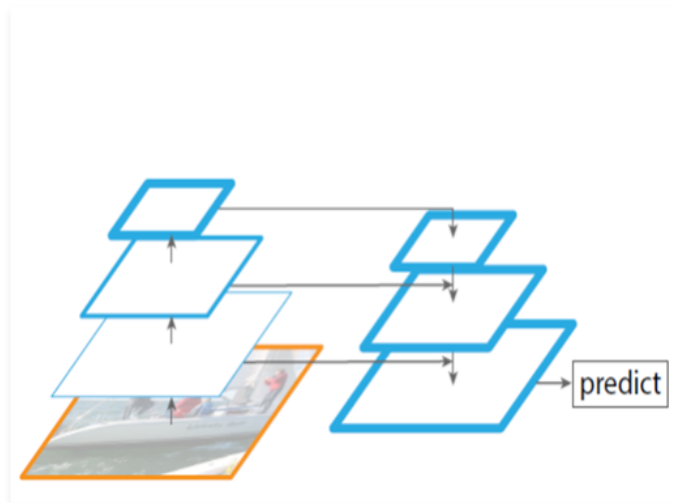


VH

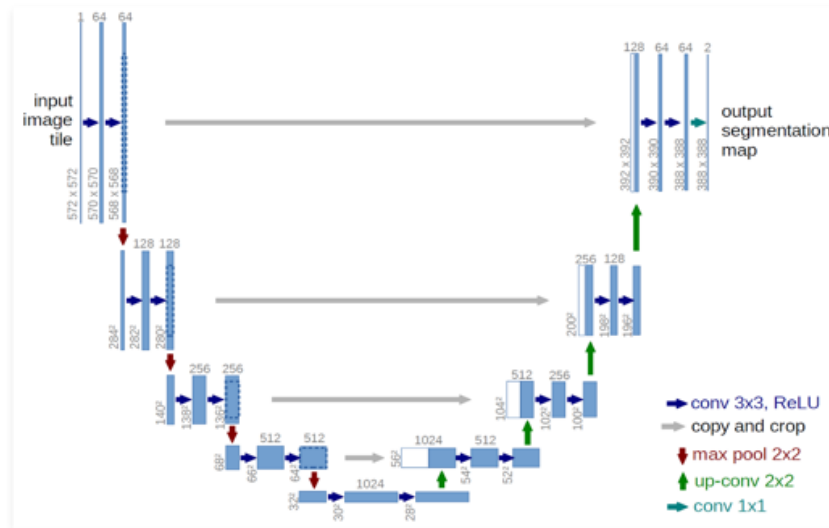


Benchmark model development

Baseline models: Vanilla FPN and U-Net (ResNet 50 encoder). random selection, 4:1 Train-Validation split



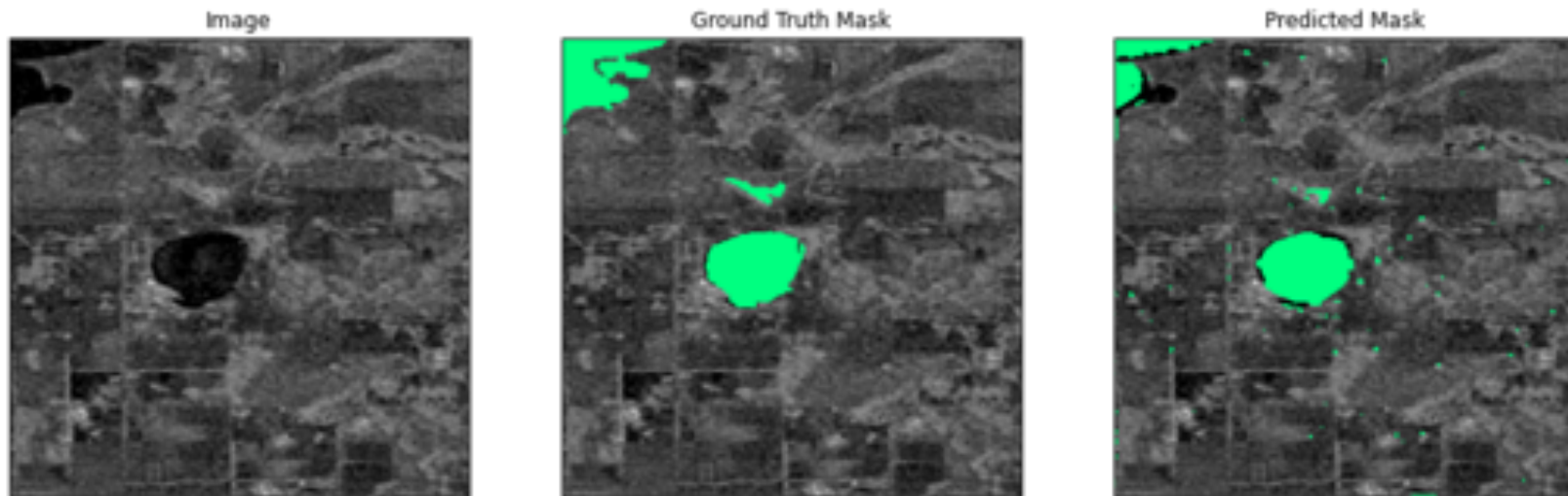
Feature Pyramid Network (Resnet 50 encoder)



U-Net (Resnet 50 encoder)

Benchmark model development

Visual Results



Sample segmentation

Leveraging citizen science for optimal model

Finding optimal modal is an exhaustive task

ML Competition in collaboration with IEEE

- 137 participants
- More than 200 submissions
- Codalab platform

- **Phase 1 (Development):** Participants are provided with training data (which includes reference data) and validation data (without reference data until phase 1 concludes) to train and validate their algorithms. Participants can submit prediction results for the validation set to the codalab competition [website](#) to get feedback on the performance from April 15 to May 14, 2021. The performance of the best submission from each account will be displayed on the leaderboard.
- **Phase 2 (Test):** Participants receive the validation set reference data for model tuning and test data set (without the corresponding reference data) to generate predictions and submit their binary classification maps in numpy array format from May 15 to June 30, 2021. After evaluation of the results, three winners will be announced on July 1, 2021.

CodaLab

[Search Competitions](#) [My Competitions](#) [Help](#)

Competition



ETCI 2021 Competition on Flood Detection

Organized by Shubhankar - Current server time: Jan. 14, 2022, 2:35 a.m. UTC

First phase

End

Development (Phase 1)

April 15, 2021, midnight UTC

Competition Ends

July 15, 2021, 11 p.m. UTC

[Learn the Details](#)

[Phases](#)

[Participate](#)

[Results](#)

[Public Submissions](#)

[Forums](#)

[Overview](#)

[Evaluation](#)

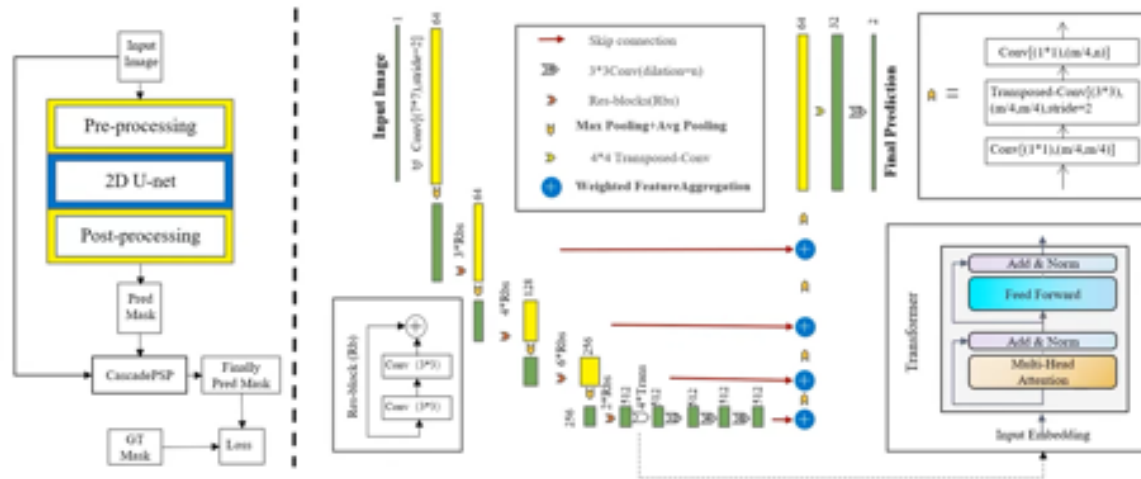
[Terms and Conditions](#)

Welcome to 2021 ETCI Flood Detection Competition

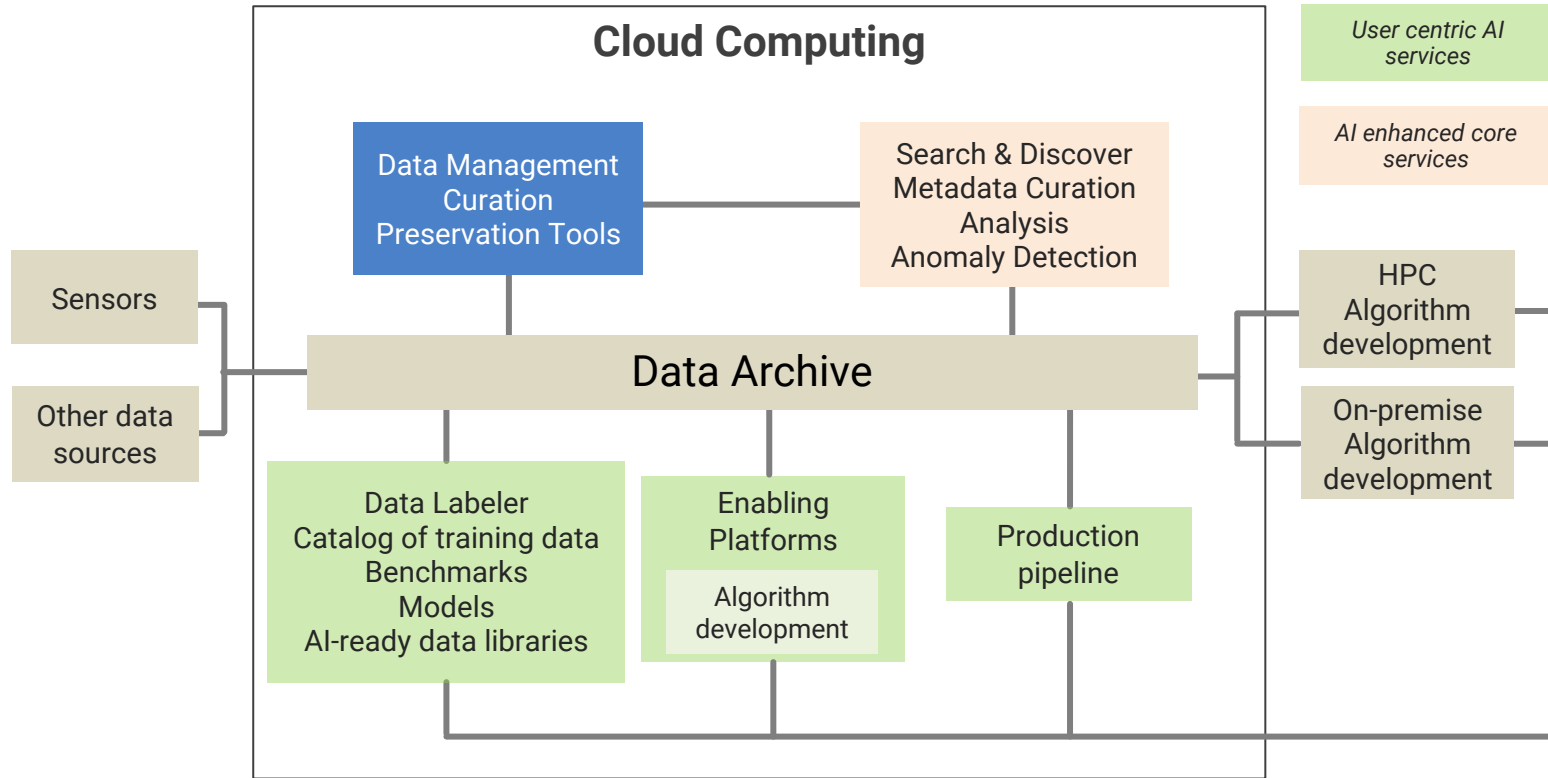
The flood event detection contest, organized by the NASA Interagency Implementation and Advanced Concepts Team in collaboration with the IEEE GRSS Earth Science Informatics Technical Committee, seeks to develop approaches to delineate open water flood areas as an effort to identify flood extent, an impactful disaster that occurs frequently throughout the world. The competition involves a supervised learning task—participants will develop algorithms to identify flood pixels after training their algorithm against a training set of synthetic aperture radar (SAR) images. Participants are required to submit binary classification maps, and performance will be evaluated using the intersection over union (IOU) score.

For details about the 2021 ETCI Flood Detection Competition: <https://nasa-impact.github.io/etci2021/>

Winning solutions



Team Arren, IOU: 0.7681



Summary

AI enhanced enterprise data systems

- AI approaches that can efficiently operate as a part of the core of large-scale systems
- Before AI can be widely used in critical enterprise data systems, we need new robust pipelines to systematically manage AI lifecycle
- A flexible architecture that allows software systems and AI algorithms to evolve to take advantage of emerging trends in hardware and software and rapid model deployment



Thank you.